Project 4

Segmenter les clients du marketplace brésilien OLIST

Clara Yaiche

OpenClassroom

2023

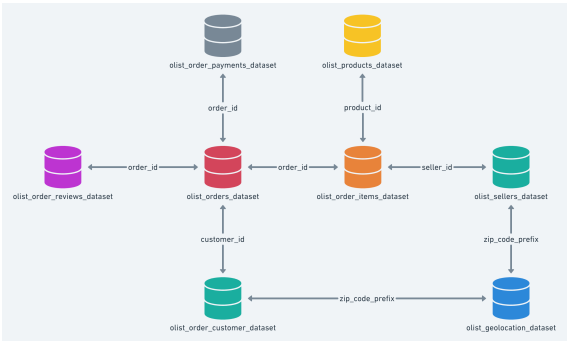
Sommaire

Les bases de données OLIST : analyse exploratoire et nettoyage

Présentation de la problématique

But : Segmenter les clients du site internet de vente en ligne OLIST. Proposer aux équipes marketing la segmentation et une temporalité de mise à jour. Il s'agit de comprendre les différents types de clients grâce à leur comportement d'achat et à leurs données personnelles.

Les données : sont séparées en plusieurs bases à fusionner. Beaucoup d'informations personnelles sur les clients sont manquantes pour des raisons de confidentialité.



Les bases de données OLIST : analyse exploratoire et nettoyage

Analyse exploratoire

Les données sont structurées de la manière suivante :
Chaque ligne correspond à un article acheté comportant des informations sur le client, son domicile, l'article, le payement ainsi que le délai et le statut de la livraison.
L'idée de mon analyse exploratoire est de répondre à certaines questions à partir des données disponibles :
- Qu'est-ce que les clients achètent le plus? À quelle époque ? Quels sont les délais de livraison ?
- D'où sont les clients sont-ils originaires, comment payent-ils leurs achats ?

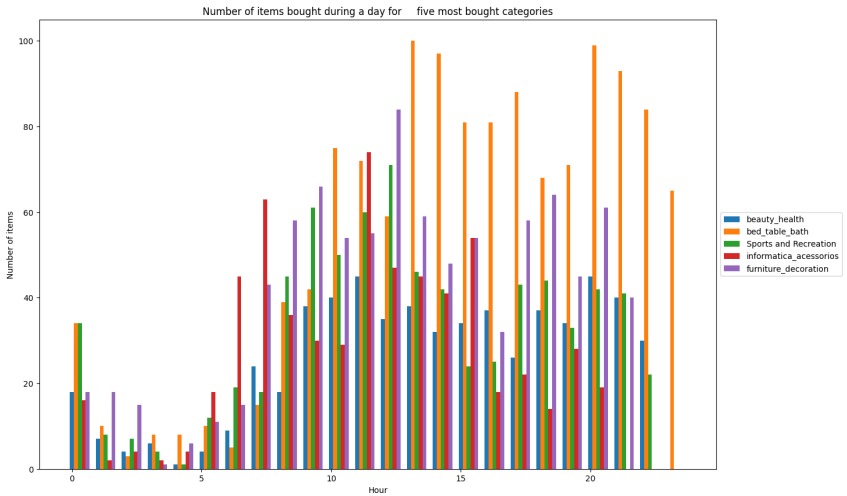
	product_category_name_translation	product_category_name	order_id	percentage	cumul_percentage
0	bed_table_bath	cama_mesa_banho	1307	0.16	0.16
1	furniture_decoration	moveis_decoracao	905	0.11	0.27
2	Sports and Recreation	esporte_lazer	752	0.09	0.36
3	informatica_acessorios	informatica_acessorios	611	0.07	0.43
4	beauty_health	beleza_saude	602	0.07	0.50
5	housewares	utilidades_domesticas	475	0.06	0.56
6	watches_gifts	relogios_presentes	329	0.04	0.60
7	tools_garden	ferramentas_jardim	306	0.04	0.64
8	fashion_bags_and_accessories	fashion_bolsas_e_acessorios	284	0.03	0.67
9	telephony	telefonia	234	0.03	0.70

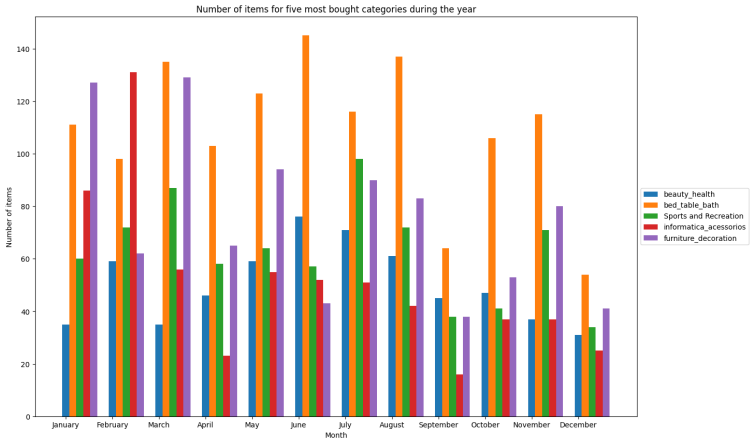
Figure: Catégories des produits les plus vendus par le site internet OLIST

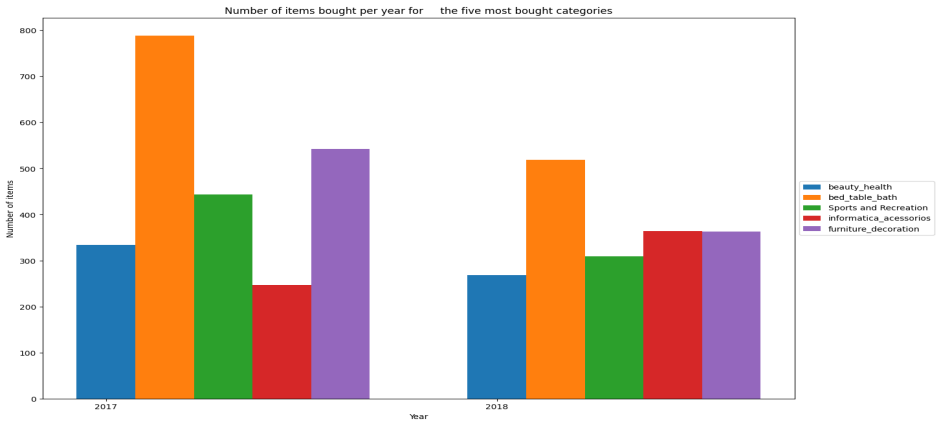
Les bases de données OLIST : analyse exploratoire et nettoyage

Analyse exploratoire : évolution des ventes

Nombre de produits totaux vendus en 2017 et 2018 par heure en mois et année. Baisse des ventes de dix pour cent entre 2017 et 2018.







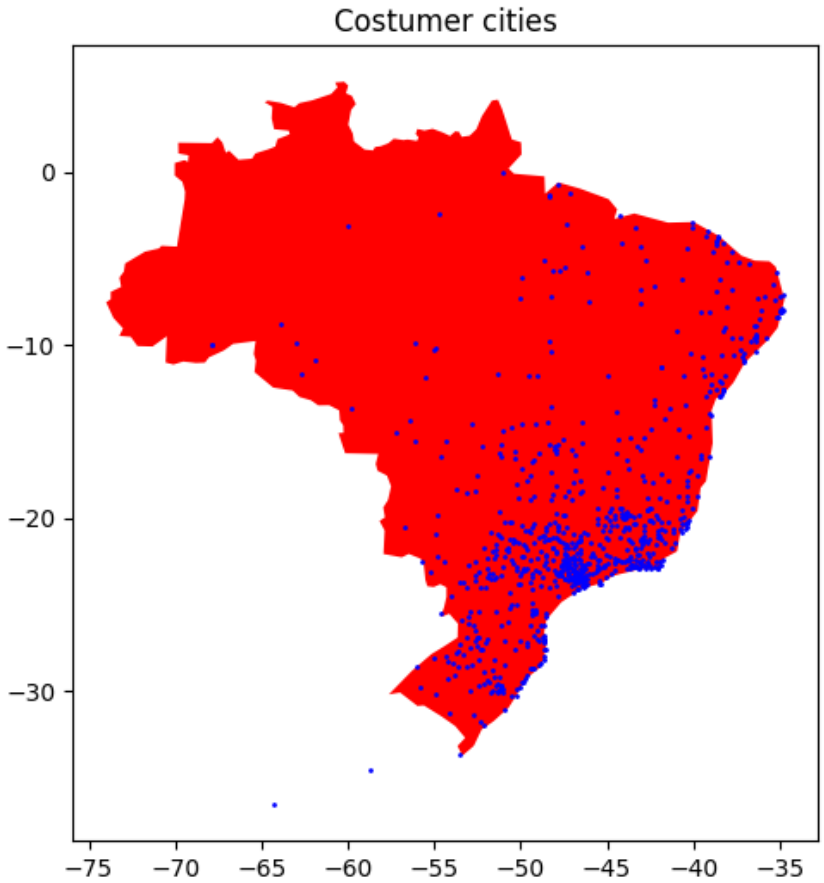
	Number of Items	Total sale price (in Brazilian Real)
2017	4288	431734
2018	3883	392357
Decrease (%)	10	10

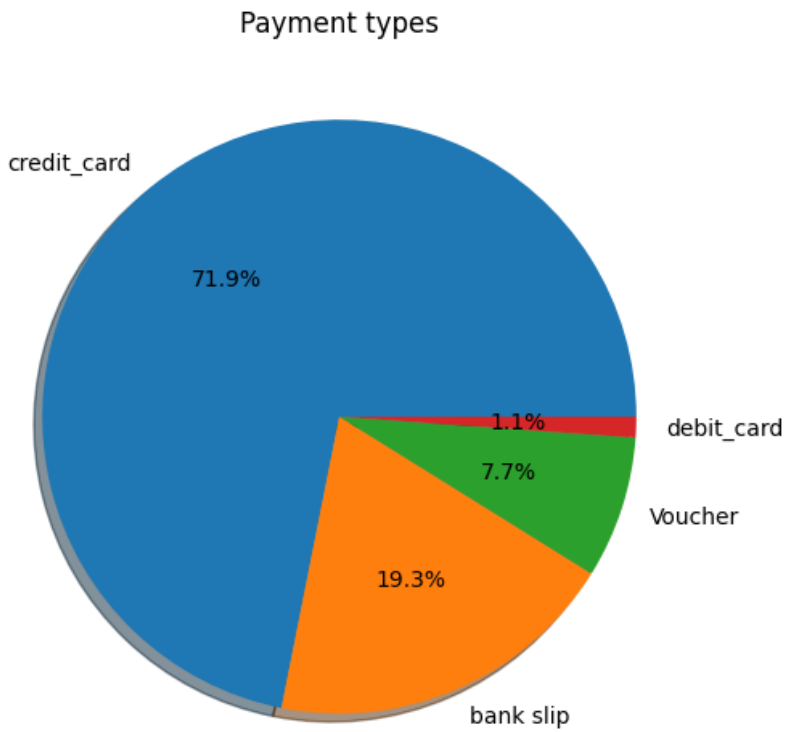
Les bases de données OLIST : analyse exploratoire et nettoyage

Analyse exploratoire : localisation et payement

La plupart des clients vivent au Brésil et payent par carte de crédit.

Le délai de livraison peut varier entre 1 et 88 jours, la médiane est de 10 jours.





Les bases de données OLIST : analyse exploratoire et nettoyage

Gestion des valeurs manquantes et atypiques

Valeurs manquantes :
- nettoyage par seuil : les colonnes plus de 50 % les individus avec plus de 70 % de valeurs manquantes.
- imputation par la moyenne
- Suppression des individus : pour les variables comportant très peu de valeurs manquantes, on supprime les individus correspondants.

Filtrage
- Par année : suppression des données de 2016 (seul 0.3% de la base)
- Par nombre de commandes : on ne garde que les clients qui ont effectué plus de deux deux commandes (qui peuvent contenir plusieurs produits), donc 3% des individus présents dans la base

Segmentation des clients du site de e-commerce

RFM (Récence, Fréquence, valeur Monétaire)

Récence : à partir de *order_purchase_timestamp*, transformer au format pandas **datetime**. Ici, on souhaite avoir un délai, on prend comme valeur référence la première date de commande :

$$\text{Récence} = \text{order\_purchase\_timestamp(Last client order)} - \text{min(order\_purchase\_timestamp)}$$

Fréquence : nombre d'achat(s)

Montant : : Somme total des achats effectués

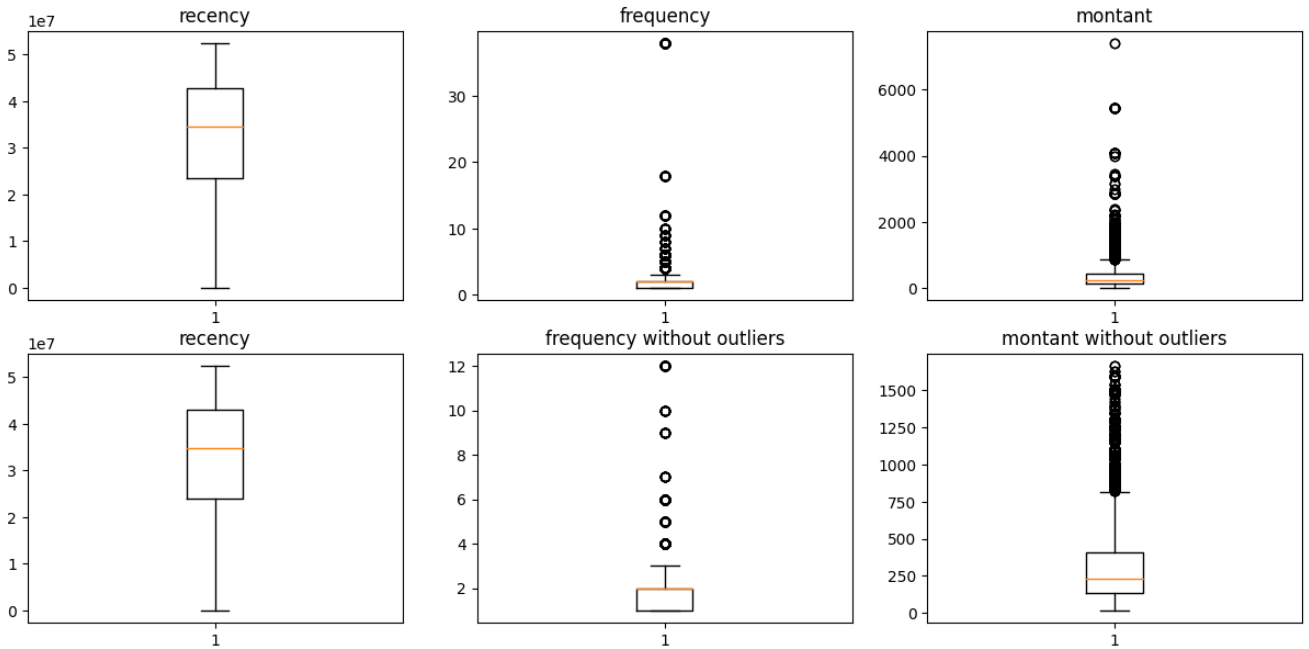
Gestion des outliers :

- **l'ecart interquartile** mais 15 % des individus sont supprimés

$$IRQ = Q_3 - Q_1$$

- **l'écart type** ici 3 % des individus sont supprimés comme les variables ne suivent pas exactement un normal qui devrait donner :

$$\mathbb{P}(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0,9973$$



Remove outliers by standard deviation

Segmentation des clients du site de e-commerce

Feature engineering : transformation et autres variables

Prise en compte d'autres critères :
- **Payement échelonné moyen** (*payment_installments*)
- **Satisfaction** : note moyenne donnée par le client (*review_score*)
- **localisation** : variable synthétique correspondant à la distance par rapport à la capitale brésilienne (nouvelle variable synthetique : *costumer_dist_from_capital*)

Outliers : suppression suivant l'écart interquartile.

Transformation des différentes variables :
- logarithmique : pour les données à très forte asymétrie à droite (pour la fréquence et la distance à la capitale).
- Transformation boxcox (pour la récence, le montant et le payement échelonné)

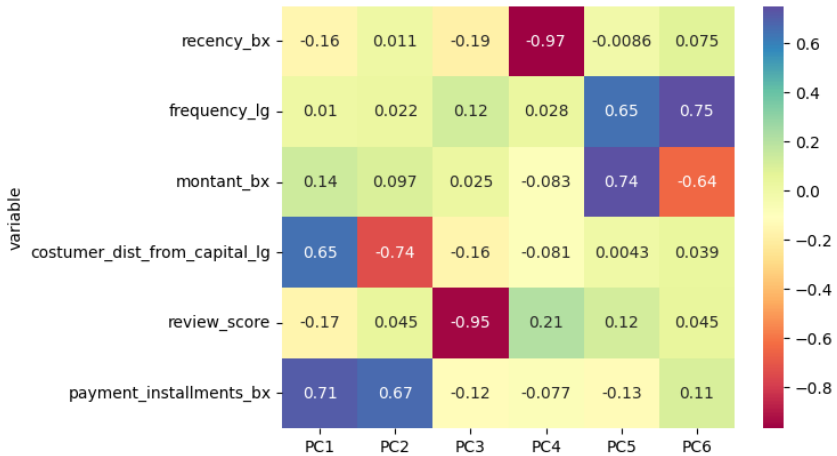
$$B(x,\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(x) & \text{si } \lambda = 0 \end{cases}$$

Mise à l'échelle : certain algorithme de clustering et surtout le K-means travaillant sur des distances, il est important de mettre à la même échelle les données. Pour cela, j'ai utilisé la class *sklearn.preprocessing.MinMaxScaler*.

Segmentation des clients du site de e-commerce

Feature engineering : transformation et autres variables

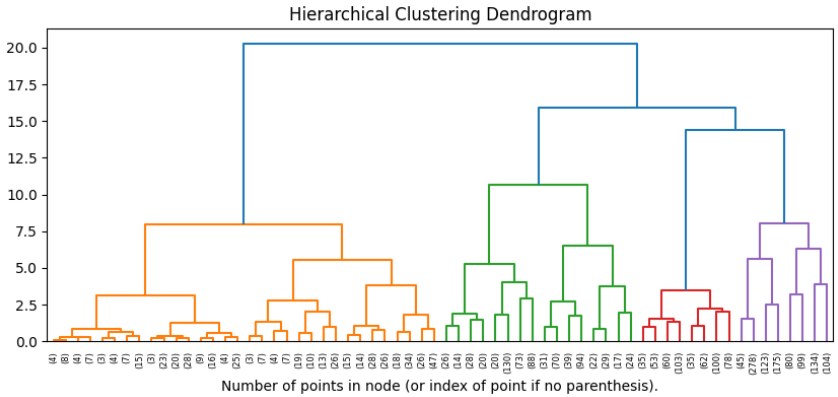
TEST différentes combinaisons :
- Uniquement RFM
- RFM et les nouvelles variables
- RFM, les nouvelles variables et la PCA

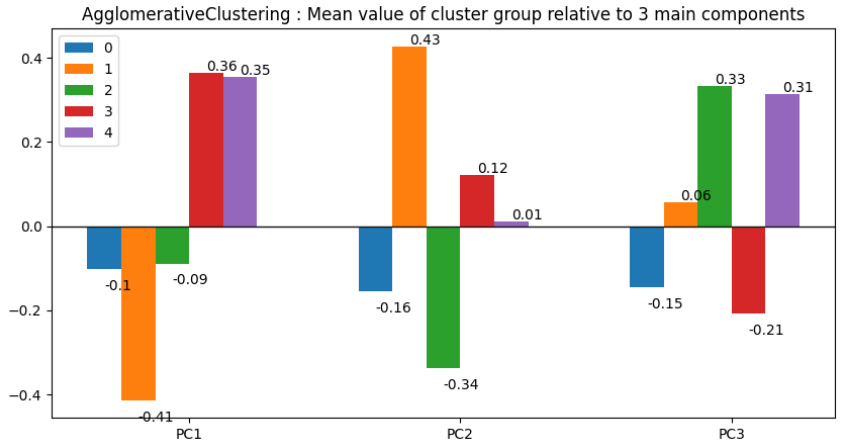


On compare ces différentes combinaisons avec les scores de silhouette et de Davies-Bouldin obtenu avec l'algorithme **K-means**.

Segmentation des clients du site de e-commerce

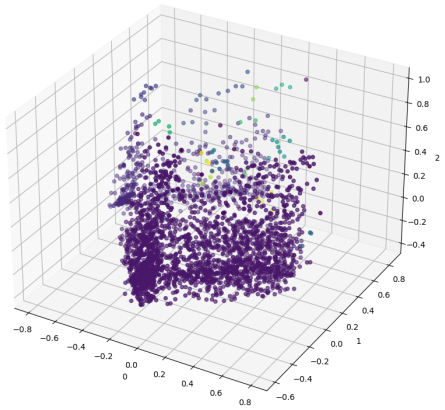
Clustering hiérarchique : agglomeratif

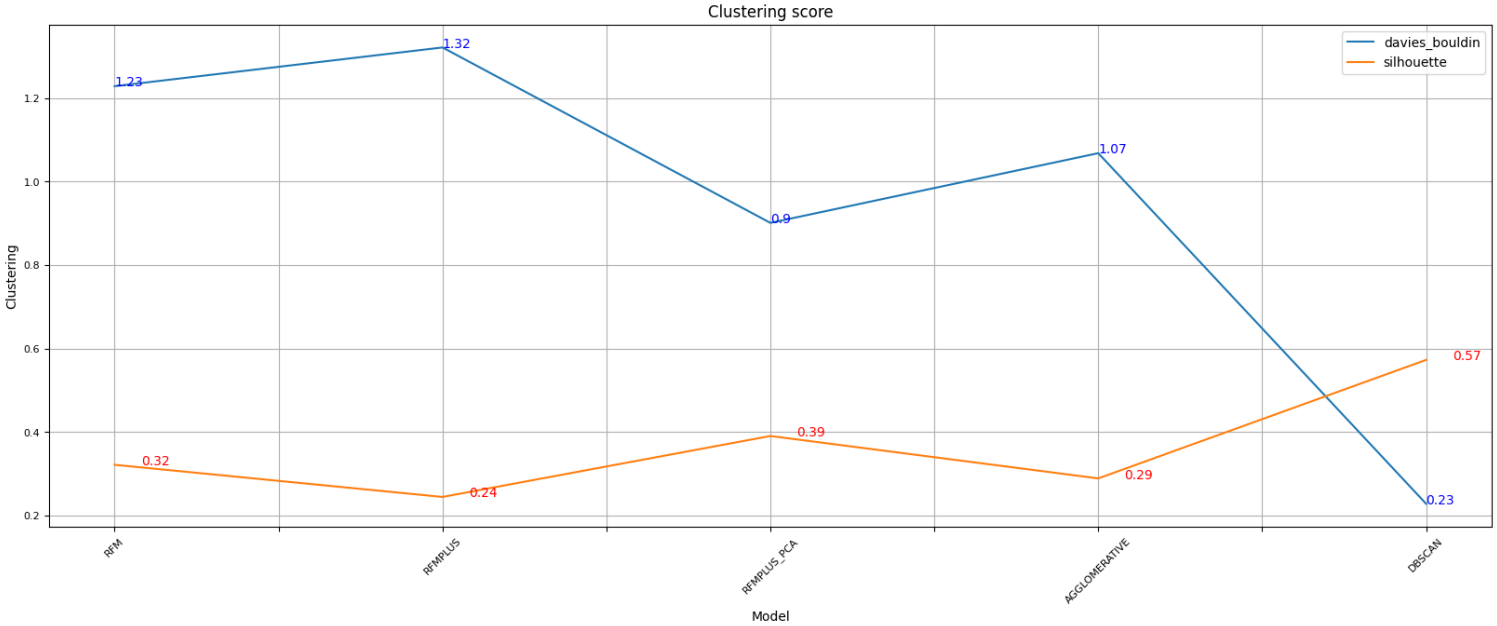




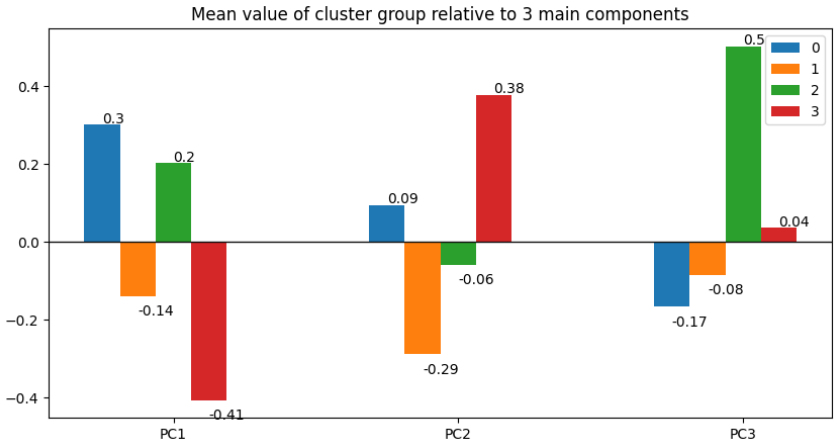
Segmentation des clients du site de e-commerce

DBSCAN and conclusion





Segmentation finale choisie : K-means (RFMPLUS et PCA)




```
1  df_pca['label'].value_counts()
✓ 0.1s

1    922
0    889
3    545
2    410
Name: label, dtype: int64
```

La compréhension des clusters après l'ACP est plus compliqué, j'ai néanmoins essayé de donner une analyse en fonction des variables portées par ces composantes :

- groupe 0 : client satisfait, payant en plusieurs fois
- groupe 1 : clients habitant proche de la capitale
- groupe 2 : clients insatisfaits et habitant loin de la capitale
- groupe 3 : clients habitant loin de la capitale

Établir le délai de maintenance

Pipeline de test

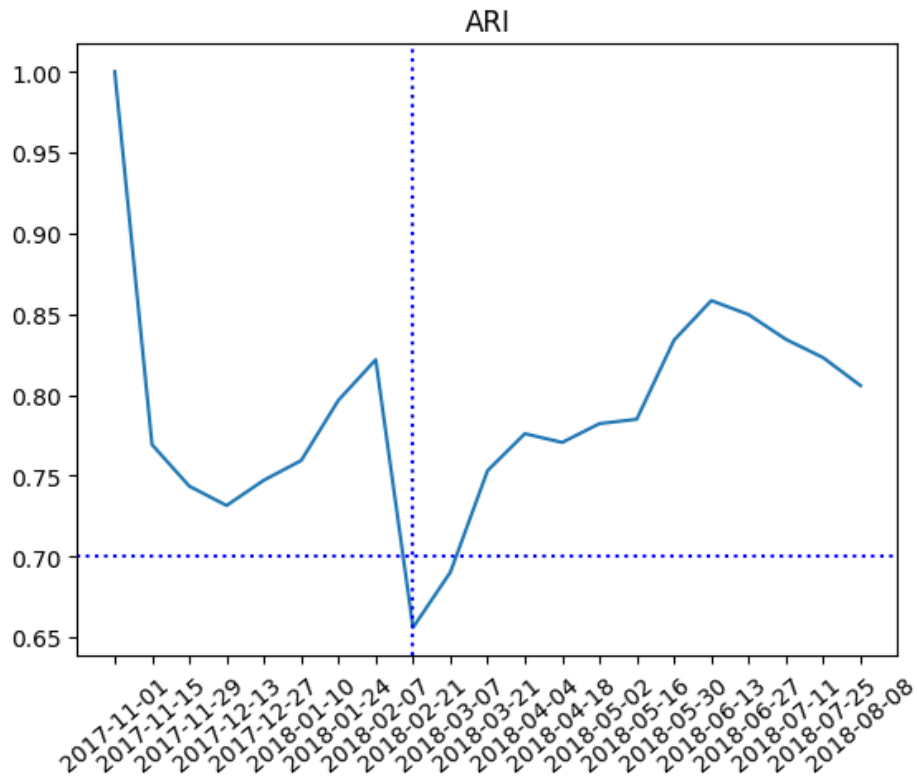
Proposition de maintenance

- basée sur l'ARI (indice Rand ajusté)
- entre novembre 2017 à août 2018.
- Considère la base pré-nettoyée : plus de valeurs manquantes ou aberrantes

Initialisation :
Cluster initial sur les commandes avant novembre 2017
Tant que la date != aout 2018 mettre à jour la base de données en tenant compte des nouvelles commandes passées par les clients. Appliquer le feature enginneering. Nouveau clustering avec K-means et calcul de l'ARI entre le clustering initial et le nouveau clustering. Ajouter 15 jours à la date courante.

Établir le délai de maintenance

Evaluation de l'ARI



Conclusion : la mise à jour de l'ARI doit être effectuée tous les quatre mois.